THE NUCLEOTIDE SEQUENCE OF A GENE ENCODING A LOW MOLECULAR WEIGHT GLUTENIN SUBUNIT FROM HEXAPLOID WHEAT

by

ERNEST GERARD PITTS

B.S., St. John's University, 1982

A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

GRADUATE BIOCHEMISTRY GROUP

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1988

Approved by:

Charles Hargrove

Major Professor

Department of Biochemistry

ACKNOWLEDGEMENTS

I would like to thank Dr. Hedgcoth for his guidance and encouragement throughout the course of the research project. The excellent advice he gave greatly expedited the completion of this thesis. My sincere thanks go to J. Antoni Rafalski for supplying the genomic clone used in this study, and to Kay Scheets for her help in planning experiments and suggestions on technical aspects of cloning and sequencing DNA. Harold Rathburn and Tino Unlap supplied valued friendship and encouragement.

I want to express my heart-felt appreciation to my wife, LuAnn, and to my family who provided much support and love throughout this endeavor.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF FIGURES AND TABLES

page

Introduction

Wheat: The Most Important Cereal Grain

Worldwide wheat production in 1986 was 535.4 million tons with a 1987 forecast production of over 508 million tons. The United States is a major wheat producer accounting for over ten percent of world production, in 1986, 56.9 million tons. Wheat is also a major product of Far East Asia with China, India, and Pakistan combining to produce almost a third of the world's wheat. In the last several years wheat has surpassed rice as the most produced and consumed cereal grain in the world (1). The protein that wheat supplies is the greatest single source in the human diet, with a significant amount of lesser quality wheat being fed to livestock (2). The reason that wheat has become so important to human civilization is two-fold, first is its unique property of forming a dough, and second is its high overall percentage of protein, from 9-17 % of the total grain by weight (3). By comparison, rice has 8.5 %, barley 11%, and corn 9.5% protein by weight (4). The protein derived from wheat is deficient in the one of the essential amino acids, lysine. This trait is directly reflected in the largest group of wheat proteins, the storage proteins, since they are typically very low in lysine. There is not, however, much study

underway to improve the quality of wheat proteins, since the increase in quality is inversely related to grain yield and may be deleterious to breadmaking properties (4). It has been concluded that the storage proteins of wheat, the gliadins and glutenins, allow wheat flour to form a dough by complexing with starch molecules in the presence of water. This complex is held together by strong hydrophobic interactions as well as intermolecular disulfide bonds (3,5).

Wheat Stroage Proteins

The wheat storage proteins are the largest group of proteins in the developing seed, a nitrogen, carbon and sulfur source found exclusively in the endosperm (5). They are found in the developing endosperm 12-15 days post-anthesis (7), and upon maturation of the seed (49 days) account for some 70-80 % of the total protein of the grain (8,9). The storage proteins of wheat are deposited in protein bodies in the endosperm after their synthesis on the rough endoplasmic reticulum (RER), their accumulation in the RER, and their subsequent cleavage from the RER to form protein bodies (10).

The Classification of Wheat Storage Proteins

The storage proteins of cereal grains have been classically grouped according to their solubility in various solvents. Osborne, in his book, The Vegetable Proteins (11), classified the storage proteins on the basis of solubility into four distinct groups, which were extracted from grain sequentially. The storage proteins were identified in the following way ; the first group, the albumins, are water-soluble, the globulins are soluble in dilute saline solutions, the prolamins are soluble in aqueous alcohols (classically 70 % ethanol), and the last group, the glutelins, are soluble in dilute acid or alkali. The storage proteins of wheat have been given the trivial names gliadin for the prolamin fraction and glutenin for the glutelin fraction. There is, however, a difficulty with this method of classification as pointed out by Payne et al. (3). The problem lies in the extractability of the proteins at each step, with some residual protein being carried over from the preceding step in the sequential extraction procedure. This is partially overcome by repeating each step of the extraction before moving on to the next solvent. It has been suggested by Kreis et al., that the storage proteins of wheat be classified according to their amino acid composition and molecular weight (9). This proposed

nomenclature somewhat simplifies the storage proteins of cereal grains with only three groups being recognized; the S-rich prolamins, the S-poor prolamins and the high molecular weight (HMW) prolamins. They also suggest altering the extraction procedure with the alcohol extraction taking place in two steps, first, as was performed classically, with alcohol alone and secondly, with alcohol in the presence of a reducing agent, such as 2-mercaptoethanol. The two resulting protein fractions have been termed prolamin-I, and prolamin-II.

The S-rich prolamins consist of what has classically been termed gliadin, and also includes an aggregated group of proteins. The gliadins are contained in the prolamin-I fraction, and the aggregated gliadin fraction is found mostly in the prolamin-II fraction, while a portion of the aggregated group is found in the prolamin-I fraction. The S-poor prolamins are contained in the prolamin-I fraction and classically were termed the omega gliadins. The HMW prolamins constitute the classical glutenin fraction. The classification scheme for the storage proteins is not ideal and will unodoubtedly undergo changes as our knowledge of them increases.

Within the aggregated group of S-rich prolamins, there are two classes that are recognized, the aggregated gliadins and the low molecular weight (LMW) glutenins. Tatham et al. contend that there is a slight difference

between these two groups in amino acid composition. In four different cultivars determined, they reported slightly more cysteine and proline and less glutamate and glutamine in the aggregated gliadins than in the LMW glutenins (12). This distinction may turn out to be erroneous as more data on this group of proteins becomes available.

Classification of Gliadins and Glutenins

The gliadins, classically defined as those grain proteins soluble in 70% ethanol, have been classified by their mobility on starch gels using aluminum lactate buffers at pH 3.1 (13). They have arbitrarily been divided into four groups designated α, β, τ and omega in descending order of mobility in starch gels. Because of the heterogeneity of various lines of wheat with respect to the gliadin proteins, electrophoresis has been used as a means of identifying different cultivars, owing to the prescence or absence of bands in the electrophoretic pattern (3,4).

The glutenins are simply designated high molecular weight (HMW) glutenin subunits or low molecular weight (LMW) glutenin subunits based on their size as deduced from polyacrylamide gels. The LMW and HMW glutenin subunits exist in aggregates stabilized by disulfide bonds

and strong hydrophobic interactions (5). The glutenins can be reduced to single polypeptide chains after treatment with 2-mercaptoethanol which reduces the disulfide bonds between the subunits. It has been observed that the glutenins impart elasticity to a bread dough while the gliadins impart extensibility to the dough. Both of these characteristics are indispensable and allow wheat flour to form a dough that can be drawn out and formed to create a loaf (5).

Wheat Prolamin Gene Loci

Common bread wheat, Triticum aestivum, has a hexaploid genome consisting of 21 pairs of chromosomes divided into three homologous genomes (designated A, B, and D) of seven chromosome pairs each. The genes that encode the wheat storage proteins have been identified (8). These genes have been assigned to nine complex loci on chromosomes 1A, 1B, and 1D, and 6A, 6B, and 6D. The genes encoding the HMW glutenin subunits are located on the long arms of chromosomes 1A, 1B, and 1D, and the genes for the LMW glutenin subunits are located on the short arms of the same chromosomes. The genes encoding the γ-gliadins, omega gliadins and a few of the β-gliadins are located on the short arms of chromosomes 1A, 1B, and 1D, and the genes encoding the α-gliadins, β-gliadins and a

few of the γ-gliadins are found on the short arms of chromosomes 6A, 6B, and 6D.

Payne et al. have studied the relationship between protein content associated with bread-making quality and allelic variants and concluded that the LMW and HMW glutenin subunits are the storage proteins most closely linked to bread-making quality (8,39).

Sequences of Wheat Storage Protein Genes

As stated earlier, the storage proteins of wheat are extremely important since they provide a major source of dietary protein for the world's population. Consequently, this family of proteins has been the topic of intensive study. For a general review, see Kreis et al. (9).

The construction of a cDNA library from wheat endosperm poly(A^+) RNA by Scheets and Rafalski (19) has provided several useful cDNA clones. By probing a wheat genomic library with a cDNA clone from this library (pW8), Rafalski et al. (19) were able to characterize the first genomic clone which encoded an α/β-gliadin. The nucleotide sequence of a cDNA clone, pW10, from this library was determined and it was found to code for a partial γ-gliadin (20). This cDNA clone was used to screen a wheat genomic library resulting in the determination of a genomic DNA sequence encoding a γ-

gliadin (21). Genomic sequences encoding HMW glutenin subunits have also been published (22,23,24), thus the LMW glutenins/aggregated gliadins remains as the only group of wheat storage proteins which has not been adequately characterized at the nucleotide level. The only sequence determined from this group is the cDNA sequence of an aggregated gliadin, B11-33 (14). A partial cDNA clone (pW12) from the library of Rafalski et al. (19) was sequenced at the nucleotide level (K. Scheets, unpublished data) and is strongly homologous to the sequence of the aggregated gliadin, B11-33 (14). The cDNA clone B11-33 was originally reported to be a γ-gliadin, but upon comparison with the reported amino terminus of a mixture of aggregated gliadins (16), B11-33 is correctly identified as an aggregated gliadin. The cDNA clone pW12 was used to probe a wheat genomic DNA library, yielding the genomic clone characterized in this study. The genomic clone sequenced in this study (LP11) shows strong homology to both pW12 and B11-33, thereby placing it in the LMW glutenin group of storage proteins.

The upstream sequence of a putative LMW glutenin gene, hereafter referred to as LMW-Colot, has been published (15), but the nucleotide sequence reported in this thesis is the first reported full-length genomic sequence encoding an LMW glutenin subunit.

Very little data has been obtained from amino acid

sequencing of the storage proteins of wheat for two reasons. The first difficulty in obtaining reliable amino acid sequence data is due to the large number and structural homology of the wheat storage proteins. Due to their structural homology the physical properties of many of the storage proteins are identical, which in turn makes their purification to homogeneity difficult. This problem is compounded by wheat having a hexaploid genome with three seperate yet related genomes, which increases the chance of having two or more proteins manifesting identical or nearly identical physical properties. The second difficulty in obtaining amino acid sequence data from the wheat storage proteins is a result of their high composition of glutamine residues. Under the acidic conditions maintained during repeated Edman degradations the exposed glutaminyl residues may undergo cyclization to pyroglutamyl residues (16). Pyroglutamic acid residues do not have a free amino group, so the peptide is effectively blocked to further cycles of Edman degradation. This is compounded by the fact that the gliadins and LMW glutenins have regions containing as many as 18 glutamine residues in tandem. The occurence of just three glutamine residues in tandem will effectively halt further cycles of Edman degradation (16). As a consequence of these difficulties, only N-terminal amino acid sequences of about 20 residues in length have been published representing the different

classes of wheat storage proteins. The difficulties associated with amino acid sequencing outline the need to sequence these genes at the nucleotide level to add to our knowledge of the storage proteins in general and the LMW glutenin subunits in particular.

Materials and Methods

Wheat Genomic Library

Wheat genomic libraries were previously incomplete because of poor cloning efficiency in the conventional lambda vectors available. Since the construction of the lambda Charon 32 vector the difficulties associated with DNA rearrangements and low cloning efficiency have been alleviated (17). The genomic library used in this study was supplied through the courtesy of Drs. M. Murray and J. Slightom of Agrigenetics Co., and was constructed by partially digesting wheat DNA with Eco RI and inserting it into the Eco RI digested lambda Charon 32 vector (18).

Subcloning a Genomic Fragment

Rafalski et al. (19), found two general size classes of poly $(A)^+$ RNA in wheat endosperm; class I corresponding to approximately 1350-1400 nucleotides in size and class II of 1550-1600 nucleotides as deduced from agarose gel electrophoresis. A cDNA clone designated pW12 hybridized to both class I and class II wheat endosperm poly $(A)^+$ RNA (19).

A LMW glutenin genomic clone was identified by screening a partial Eco RI wheat genomic library with the

cDNA clone pW12. A strong positive signal was obtained for the genomic clone LP11, which released an 8 kb non-hybridizing fragment, and a 5.2 kb hybridizing fragment upon digestion with Eco RI. This 5.2 kb EcoRI fragment was subcloned in both orinetations into M13mp19, and designated MW11 and MW11R. Southern hybridization with pW12 showed that a 1.8 kb Sph I/Eco RI fragment contained all of the hybridizing insert. Subcloning of MW11 by digestion with Sph I and recircularization excised a non-hybridizing portion of the original 5.2 kb insert, leaving a 1.8 kb insert, which facilitated subcloning and manipulation. This subclone, designated MW11/Sph, was used for further subcloning strategies.

Generating Deletion Subclones

The strategy of constructing deletion subclones (15) for sequencing employed an exonuclease which removed nucleotides from the insert at a linear rate. By removing aliquots from the reaction at various time points, the population of molecules was deleted to a varied extent yielding a library of clones with differing amounts of insert remaining. These deleted molecules were then ligated, and a series of overlapping clones was obtained which covered the entire insert.

One of the comercially available oligonucleotide primers was hybridized to the single stranded DNA (ssDNA) molecules containing inserts. Hydridizing the primer regenerated a restriction site, which upon digestion left the 3' end of the insert exposed. Deletion subclones were then generated utilizing the 3' to 5' exonuclease activity of T4 DNA polymerase. The shortened insert-containing ssDNA molecules were homonucleotide tailed using terminal transferase (TdT), the oligonucleotide primer was re-annealed, and ligation was promoted by T4 DNA ligase. The ligated DNA was transformed into competent cells of E.coli strain JM109, and the transformants containg deletion subclones were recovered.

Preparation of ssDNA

Recombinant DNA was transformed into competent cells of E. coli JM109, which were plated out on 1.5% L-agar plates and grown overnight. Plaques were transferred to 5 ml cultures of JM109 in L-broth at an OD_{590} of 0.03-0.06 using 6 inch wooden applicator sticks. The cultures were grown for 7-9 hr at 37°C with shaking. Cultures were centrifuged to pellet the cells. The phage containing supernatant was removed, and the phage precipitated by the addition of 1/5 vol. of 24% polyethylene glycol-6000 in 3 M NaCl. After 45 minutes on ice or overnight at 4°C, the

phage were pelleted by centrifugation. The pellets were allowed to drain inverted for 1 hr and the excess supernatant was swabbed away. Phage were then resuspended in 20 mM Tris-Cl, pH 7.5, and the protein coat was removed by the addition of 1 vol. of phenol. Aqueous and organic phases were separated by centrifugation after which two to three phenol:chloroform (1:1) extractions were performed on the aqueous phase, followed by a single chloroform extraction to remove any remaining phenol. The ssDNA was precipitated by centrifugation after the addition of ammonium acetate to 0.25 M and 2 vol. of absolute ethanol and incubation on ice for 45 min. The DNA pellet was washed with 70% ethanol, dried in a vacuum dessicator, and dissolved in 10 mM Tris-Cl, pH 8.0. The typical yield was 1-4 ug of ssDNA per milliter of culture.

Sequencing of ssDNA by the Dideoxy Chain Termination Method

Sequencing was performed using ssDNA as a template by a modification (Bethesda Research Laboratories) of the procedure of Sanger et al. (26). Primer was annealed in a 500 ul microfuge tube to the ssDNA in a volume of 12.5 ul containing 0.5 ug ssDNA, 62 ng 17-mer sequencing primer, 8 uM Tris-Cl, pH 7.5, and 8 uM $MgCl_2$. The

microfuge tube was incubated for 5 min in a water-filled 13x100 mm test tube in a 90°C water bath. The water-filled test tube containing the microfuge tube was removed from the bath and allowed to cool slowly to room temperature. The DNA was stored for a few days at -20°C or used immediately in the sequencing reactions. The following was added to the ssDNA/primer hybrid: 1 ul of 0.1 M DTT, 1 ul of [α-^{32}P]dATP (400-1000 Ci/mmol) and 0.5 ul (5u/ul) of Klenow fragment of DNA polymerase I. The contents of the tube was mixed and three ul was transferred to each of 4 tubes labled A, C, G, and T, which contained 1 ul of each of the corresponding deoxynucleotide and dideoxynucleotide mix. After 15 min at room temperature, 1 ul of 0.5 mM dATP was added to each reaction as a chase. After an additional 15 min at room temperature, the reactions were stopped by adding 10 ul of formamide loading solution (95% deionized formamide, 10 mM EDTA, pH 8.0, 0.1% each XCFF and BPB) to each reaction tube. The tubes were heated to 90°C in a heat block for 2 min and quick cooled on ice to ensure denaturation. Samples were centrifuged for 5 seconds to pellet any insoluble material before loading onto a sequencing gel. The final concentrations of the deoxynucleotides and dideoxynucleotides before executing the cold dATP chase were the following.

Base Specific Reaction	dNTP (uM)	ddNTP (uM)
A	0.6	10
C	3.75	100
G	3.75	100
T	3.25	560

Restriction Digests of DNA

Restriction enzymes were obtained from NEB, BRL, and IBI. Digestions were done using the following conditions.

Bam HI, Nde I, Sph I, Sal I
- 10 mM Tris-Cl, pH 7.5
- 10 mM $MgCl_2$
- 0.1 mg/ml gelatin
- 150 mM NaCl
- 6 mM 2-mercaptoethanol

Pst I
- 10 mM Tris-Cl, pH 7.5
- 10 mM $MgCl_2$
- 0.1 mg/ml gelatin
- 60 mM NaCl
- 6 mM 2-mercaptoethanol

Hind III	same as Pst I, without 2-mercaptoethanol
Hpa I, Sma I	10 mM Tris-Cl, pH 7.5 10 mM $MgCl_2$ 0.1 mg/ml gelatin 20 mM KCl 6 mM 2mercaptoethanol
Mbo II	10 mM Tris-Cl pH 7.5 10 mM $MgCl_2$ 0.1 mg/ml gelatin 6 mM KCl 6 mM 2-mercaptoethanol
EcoRI	70 mM Tris-Cl pH 7.5 5 mM $MgCl_2$ 0.1 mg/ml gelatin 50 mM NaCl

All digestions were carried out in a 37°C water bath, with the exception of Sma I digestions, which were incubated at 30°C.

Southern Blotting

DNA fragments were transferred from 0.7% agarose minigels onto nitrocellulose in 4 hr using a modification (27) of Southern's method (28).

Labelling of DNA by Nick Translation

Nick translation of pW12 was done in a 10 ul volume with the following concentrations: 50 ug/ml DNA, 20 uM each dGTP, dCTP, and dTTP, 0.4 uM [α-^{32}P]dATP (600-800 Ci/mmol), 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 0.1 mM DTT, 0.1 mg/ml gelatin, 8 ng/ml DNase I, and 3.75 units of DNA polymerase I (27). The solution was incubated for 1 hr at 16^{o} C, diluted to 0.1 ml using 10 mM Tris-Cl pH 8.0, 1mM EDTA, and passed over a 1 ml Sephadex G-50 quick-spin column to remove any unincorporated nucleotide monomers.

Hybridization to DNA bound to Nitrocellulose

DNA hybridizations were performed using a modification (27) of Southern's procedure (28). The baked nitrocellulose filter was wetted in 6x saline sodium citrate (SSC). Prehybridization was in 0.2 ml/cm^2 of a solution containing 2x saline sodium phosphate EDTA (SSPE), 5x Denhardt's solution, 0.1% sodium dodecyl

sulfate (SDS), and 0.1 mg/ml denatured sheared salmon sperm DNA for 4 hr at 65°C. The solution was removed and replaced with 50 ul/cm^2 of hybridization solution (same concentrations as for the prehybridization solution with the addition of 0.1 ug probe DNA at 10^7-10^8 cpm/ug). Hybridization was for 4 hr at 65°C with occasional agitation. The filter was washed for 15 min at room temperature in 2xSSC with 0.1% SDS, twice for 30 min at 65°C in 0.5xSSC with 0.1% SDS, and finally for 15 min at 65°C in 0.2xSSC with 0.1% SDS. The filter was sealed in a plastic bag and exposed to Kodak XAR-5 film for 12-24 hr at -70°C with an intensifying screen.

Cloning into M13mp18

The 1.8 kb Sph/Eco fragment originally cloned into M13mp19 was also force cloned for the opposite orientation into M13mp18 in order to sequence the complementary DNA strand. A sequential digest of MW11/Sph replicative form (RF) DNA with Sph I and Eco RI excised the 1.8 kb fragment of interest. M13mp18 RF DNA was also sequentially digested with Sph I and Eco RI, dephosphorylated, and the insert and vector ligated. (See "Ligation of DNA", below.)

Recovery of DNA Fragments from Agarose Gels

Digested DNA was separated electrophoretically on 0.7% agarose/TBE minigels, and the DNA fragment of interest indirectly visualized using UV irradiation, so as not to damage the DNA. The portion of the gel containing the insert was placed in the well of the electroeluter filled with TBE (89 mM Tris-borate, 89 mM boric acid, 2 mM EDTA) and 0.1 ml 4 M NaCl/bromphenol blue was used as a salt cushion. Electrophoresis was for 30 to 45 min at 100 volts. The DNA-containing buffer was removed from the apparatus, the DNA precipitated with 2 vol. of ethanol, and recovered by centrifugation at room temperature. The DNA pellet was washed with 70% ethanol, re-centrifuged, dried under vacuum and resuspended in TE (10 mM Tris-Cl, pH 8.0, 1 mM EDTA). The DNA concentration was determined by comparison to known standards on an agarose gel containing ethidium bromide.

Dephosphorylation of Vector

M13mp18 RF DNA (2.5 ug) was sequentially digested with Sph I and Eco RI to completion. The vector was then dephosphorylated using the procedure of Maniatis et al. (27). Calf intestinal alkaline phosphatase (CIP) in an ammonium sulfate suspension (32 units) was centrifuged for

1 min at room temperature. The supernatant was carefully removed and discarded. The CIP pellet was resuspended in 10 ul of sterile deionized water. Five microliters of the CIP solution was added to the vector DNA and the concentration adjusted to 50 mM Tris-Cl, pH 8.0, 1mM $MgCl_2$, 0.1 mM $ZnCl_2$, and 1 mM spermidine in a final volume of 0.1 ml. The reaction was incubated for 15 min at 37°C, followed by 15 min at 56°C. The remaining 5 ul (16 units) of CIP solution was added to the reaction tube and incubated as before. The concentration was adjusted to include 1% SDS, 10 mM Tris-Cl, pH 8.0, 0.1 M NaCl, and 1 mM EDTA, and the solution was heated to 70°C for 15 min to inactivate the CIP. The DNA was ethanol precipitated in the prescence of 2M ammonium acetate, recovered by centrifugation, washed with 70% ethanol and re-centrifuged. The pellet was resuspended in a small volume of TE.

Generation of Blunt Ended Fragments

Restricted DNA fragments with either 3' or 5' overhangs were converted to blunt ends in the following manner (27). Approximately 200 ng of digested DNA was adjusted to the following concentration: 33 mM Tris-acetate, pH 7.9, 66 mM potassium acetate, 1 mM magnesium acetate, and 0.2 mM each of dATP, dGTP, dCTP, and dTTP,

and 3 units of T_4 DNA polymerase were added. T_4 DNA polymerase was used because it has both a 3' exonuclease and a 5' to 3' polymerase activity so that both 3' and 5' overhangs could be converted to blunt ends simultaneously. The reaction was incubated at 37° C for 5 min, followed by a phenol/chloroform extraction, ethanol precipitation, wash with 70% ethanol, and vacuum drying. The DNA pellets were resuspended and used immediately in ligation reactions.

Ligation of DNA

For cohesive end ligations (29), 100 ng of insert DNA and 200 ng of vector DNA were coprecipitated, and resuspended in water. Ligase buffer was added to adjust the concentration to 50 mM Tris-Cl, pH 7.5, 10 mM $MgCl_2$, 10 mM DTT, 1 mM spermidine, 1 mM ATP, and 0.1 mg/ml gelatin, and 3 Weiss units of T_4 DNA ligase were added. The reaction was incubated for 1-2 h at 15° C.
The substrates for intramolecular ligation of DNA were blunt ended insert-vector molecules. Intramolecular ligations (30), were done using approximately 200 ng of DNA in a 30 ul reaction volume with the following concentrations: 25 mM Tris-Cl, pH 7.5, 5 mM $MgCl_2$, 0.1 mM ATP, and 0.06 Weiss units of T_4 DNA ligase. The reaction was incubated at 25° C for 1 hr.

Plasmid and Replicative Form DNA Preparation

Preparations of plasmid DNA and M13 replicative form (RF) DNA were done according to Maniatis et al. (27).

Identification of Cloned DNA

Insert DNA cloned into the M13 phage was identified by the methods of Messing (36). Direct gel electrophoresis was used to identify clones for sequencing, and the C-test was used to identify clones of opposite orientations.

Bacterial Transformation

Escherichia coli JM109 was the host strain for transformation. An overnight culture was prepared by transferring a colony from an M9 agar plate to 5 ml of M9 media. Approximately 15 ml of L-broth containing 20 mM $MgCl_2$ was inoculated with 250 ul of JM109 overnight culture and allowed to grow until an OD_{590} of 0.6-0.7 was detected. The cells were pelleted and incubated on ice for 10 min, and then resuspended in 1/5 vol. of transformation buffer: 10 mM Na-MES, pH 6.3, 0.01 mM $CaCl_2$, 3.6 uM hexaminecobalt (III) chloride, 45 uM $MnCl_2$,

and 0.1 mM RbCl (32). After 10 min on ice, the cells were again pelleted and resuspended in 1/12 vol. of transformation buffer. Seven microliters of dimethyl sulfoxide was added for each 200 ul of bacterial suspension, then the cells were incubated on ice for 5 min. Seven microliters of 0.75 M 2-mercaptoethanol was added for each 200 ul of bacterial suspension and the cells incubated on ice for 10 min. The addition of dimethyl sulfoxide was repeated and the cells were incubated another 5 min on ice. The DNA to be transformed was added to a polypropylene test tube on ice, and two hundred microliters of competent cells were added. The tube was vortexed and incubated on ice for 30 min. The cells were heat shocked by placing the tube in a 42^{o}C water bath for 90 sec, followed immediately by a 2 min incubation on ice. Three milliliters of 0.6% soft agar and 270 ul of log phase JM109 mix were added to the cells. The mixture was vortexed and poured over 1.5% B-plates (1% tryptone, 0.8% NaCl, 1.5% bacto agar, and 0.001% thiamine-HCl). The log phase JM109 mix was prepared by adjusting log phase cells in L-broth to 0.01% thiamine HCl, 0.75 mg/ml isopropylthiogalactpyranoside (IPTG), and 4 mg/ml 5-bromo-4-chloro-3-indolyl-β-D galactoside (X-gal). The plates were incubated overnight at 37^{o}C. Plaques were picked after 8-12 hours, with an observed transformation rate of 1-7 x 10^{5} plaques/ug ssDNA.

Agarose Gel Electrophoresis

Agarose gel electrophoresis was performed according to the general guidelines found in Molecular Cloning (27). A tris-acetate-EDTA (TAE) buffer system was used for general work while a tris-borate-EDTA (TBE) buffer was used for some of the ssDNA work (25).

Polyacrylamide Gel Electrophoresis

Six percent polyacrylamide gels for sequencing DNA were prepared by standard techniques using a 20:1, acrylamide: N,N'-methylene-bis-acrylamide ratio.

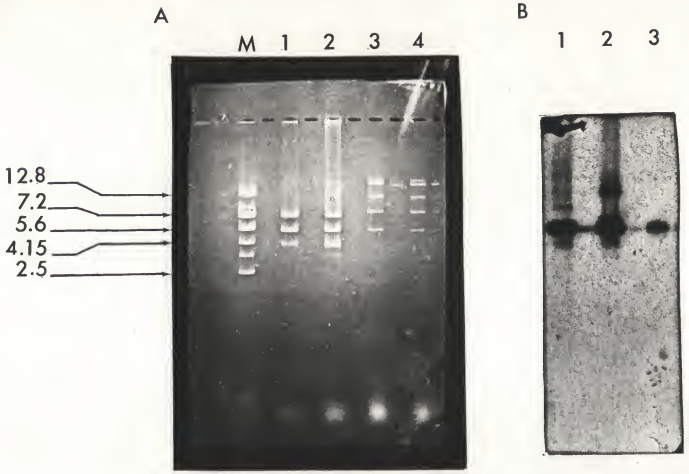
Results

Experimental Design

As described earlier, a lambda Charon 32 clone, LP11, from a wheat genomic DNA library strongly hybridized to the cDNA clone pW12 (data not shown). The 5.2 kb hybridizing fragment was isolated and cloned into the Eco RI site of M13mp19 and designated MW11. In order to demonstrate that the correct fragment had been cloned from the lambda phage into the M13 phage, the Eco RI digested lambda clone LP11 and the Eco RI digested M13 clone MW11 were electrophoresed side by side on an agarose gel (Fig 1A), blotted to nitrocellulose and probed with pW12 (Fig.1B). The 5.2 kb Eco RI fragments from LP11, MW11, and MW11R hybridized to pW12 thereby confirming the identity of both MW11 and MW11R.

Since both ends of the 5.2 kb insert had Eco RI compatible overhangs, neither orientation of the insert was favored to occur over the other, and in fact both orientations were isolated. After identifying that both orientations had been isolated by use of the complementation test and direct gel electrophoresis (36), it was observed that both orientations of the insert were unstable. It appeared that deletions and possibly rearrangements were occuring because the restriction data

Figure 1. Electrophoresis of Eco RI digests of MW11, MW11R, and LP11. Lane M: linear double stranded DNA markers (the size of these markers is shown in kb to the left of the gel). Lane 1: an Eco RI digest of MW11. Lane 2: an Eco RI digest of MW11R. Lane 3: an Eco RI digest of LP11. Lane 4: LP11 digested with a different lot of the Eco RI enzyme. Electrophoresis was at 5.5 V/cm for 2 hr. The gel outside of lanes 1 and 3 was trimmed away, the DNA was blotted to nitrocellulose and probed with pW12 (see Materials and Methods for protocol). The membrane was autoradiographed with an intensifying screen for 14 hr. B. The resulting autoradiograph on the right shows that the 5.2 kb fragment from each of the clones contains a sequence homologous to pW12.


A
M 1 2 3 4
12.8
7.2
5.6
4.15
2.5
B
1 2 3

were variable. Since both MW11 and MW11R were unstable, we decided to delete a portion of the insert, reasoning that either a smaller insert would be more stable or the region responsible for the instability might be deleted. Before carrying out this strategy it was requisite that the location of the gene be determined. To do this we digested MW11 with various restriction endonucleases, separated the DNA on an agarose gel, blotted the DNA to nitrocellulose and probed the DNA with nick translated pW12. We used the results from Fig. 2A along with results from other gels (data not shown) to construct a restriction map of the 5.2 kb Eco RI insert (Fig. 3). We were able to deduce from the autoradiograph in Fig. 2B that the gene was contained on a 1.8 kb Sph I/Eco RI fragment. A 1.8 kb Hind III fragment from a digestion of the 5.2 kb insert hybridized to pW12 while a 2.5 Sph I fragment from the 5.2 kb insert did not hybridize to pW12.

This information in turn allowed us to remove the non-hybridizing portion of MW11 by a simple strategy of digesting with Sph I and ligating to recircularize and create MW11/Sph (Fig. 4). This restriction-deletion strategy was not possible in the reverse orientation, due to the location of the gene in MW11R; therefore deletion subcloning was attempted using the method of Dale et al. (25). All of the deletion subclones isolated from MW11R had the insert totally deleted. This method was repeated

Figure 2A. Restriction data and hybridization with pW12. A 0.7% agarose minigel with 0.5 ug/ml ethidium bromide was electrophoresed at 7.8 V/cm for 3 hr. (M) designates marker lanes and the numbers to the left indicate the size of double stranded marker DNA in kilobases. The lanes marked 1-7 contain restriction digested MW11. 1: Hind III, 2: Sph I, 3: Pst I, 4: Sal I, 5: Bam HI, 6: Sma I, 7: Eco RI. The gel was trimmed to remove the marker lanes, blotted to nitrocellulose, and hybridized to nick translated pW12.

B. The resulting autoradiogram is shown to the right of the agarose gel and was exposed for 13.5 hr.

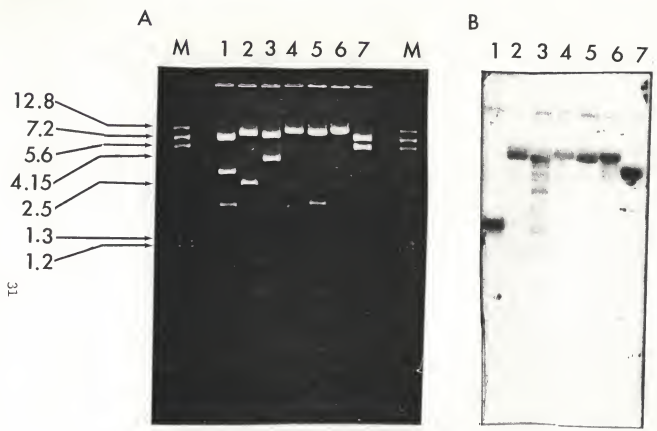
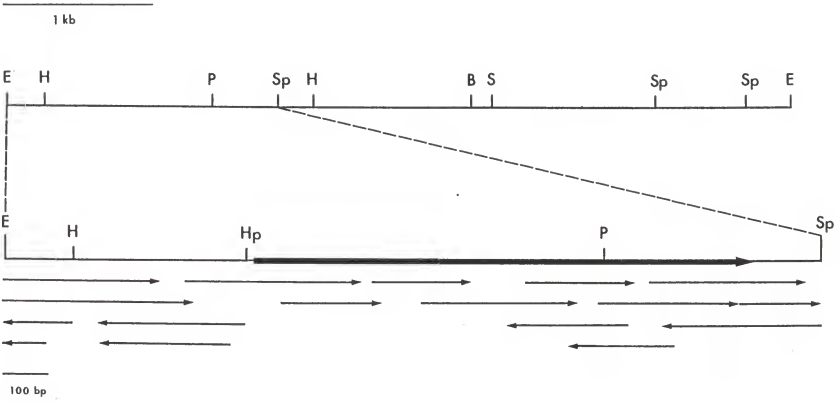

A
M 1 2 3 4 5 6 7 M
12.8
7.2
5.6
4.15
2.5
1.3
1.2
B
1 2 3 4 5 6 7

Figure 3. Restriction map of the 5.2 kb MW11 insert and sequencing strategy. The restriction map was determined experimentally using the data from the agarose gel described in figure 2 along with other agarose gels not shown. The only restriction site that was placed as a result of sequence data is the Hpa I site just upstream of the coding sequence. The bold arrow superimposed on the expanded portion (Eco RI/Sph I fragment) shows the location of the coding sequence of the LMW glutenin gene L11. The arrows below the 1.8 kb Eco RI/Sph I fragment show the sequencing strategy which employed deletion subclones exclusively (aside from the parent clones that were sequenced). The map is drawn to scale.


1 kb
E
H
P
Sp
H
B
S
Sp
Sp
E
E
H
Hp
P
Sp
100 bp

twice with the same result, thus experimental error was highly unlikely. Apparently the insert DNA was unstable in this orientation.

MW11/Sph became the substrate for deletion subcloning using the method of Dale et al. (25) and restriction-deletion subcloning strategies. Eighty-two deletion subclones were isolated with eight clones overlapping to yield the entire 1.8 kb insert. A Pst I/Eco RI subclone was also constructed from MW11/Sph (Fig.4) and its sequence confirmed that of the deletion subclones with which it overlapped. Since we already knew the location of the gene, we were able to clone the gene in the reverse orientation by excising the 1.8 kb Eco RI/Sph I fragment containing the gene and cloning it into M13mp18. This allowed the complementary strand to be sequenced. This subclone, designated MW11/SphR, underwent deletion subcloning (25) and restriction-deletion subcloning. In excess of 100 deletion subclones were screened and many of these were sequenced. The sequence data derived from these subclones was clustered on either end of the insert, leaving a portion of the insert unconfirmed by complementary strand sequence data. This difficulty was partially overcome by constructing two restriction-deletion subclones from MW11/SphR; one employing a Pst I/Sph I deletion and the other a Hpa I/Sph I deletion (Fig.5). After sequencing these subclones, the gap in the

Figure 4. Constructions of clones in M13mp19. MW11 is the 5.2 kb Eco RI hybridizing fragment from the lambda clone LP11 cloned into the multiple cloning site of M13mp19. The LMW glutenin gene is boxed and labled L11, with the arrow inside the box showing its orientation. The broad arrow adjacent to the 6230 Eco RI site shows the hybridization site for the M13 universal sequencing primer and the direction of primer extension.

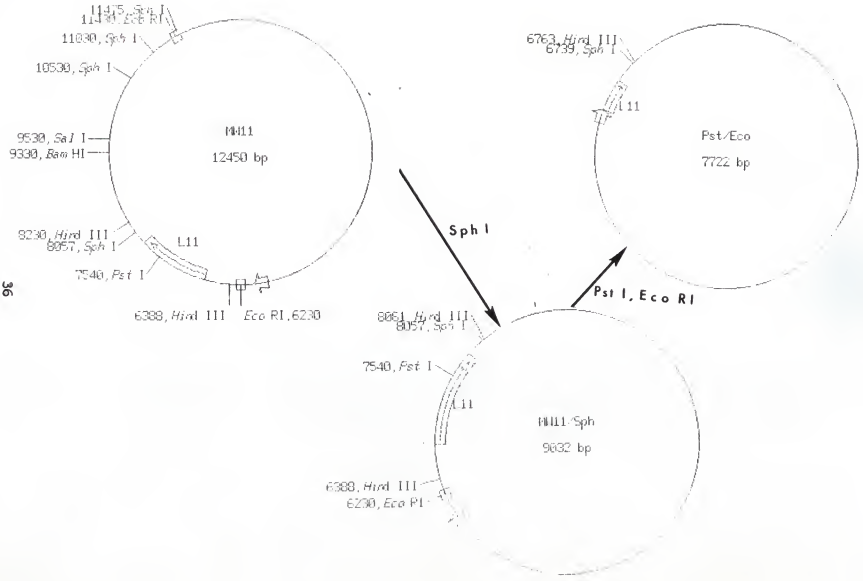
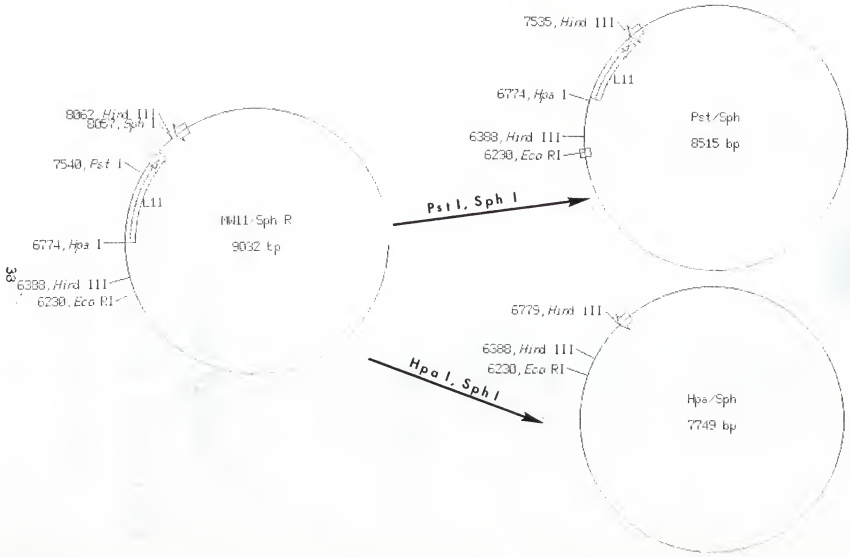

11475, Sph I
11430, Eco RI
11030, Sph I
10530, Sph I
9530, Sal I
9330, Bam HI
8230, Hind III
8057, Sph I
7540, Pst I
6388, Hind III
Eco RI, 6230
MM11
12450 bp
L11
6763, Hind III
6739, Sph I
Pst/Eco
7722 bp
Sph I
Pst I, Eco RI
8061, Hind III
8057, Sph I
7540, Pst I
L11
MM11 Sph
9632 bp
6388, Hind III
6230, Eco RI

Figure 5. Construction of clones in M13mp18. MW11/SphR is the 1.8 kb fragment cloned into the multiple cloning site of M13mp18. The gene is boxed and labled L11 as in Fig. 4. The broad arrow adjacent to the 8062 Hind III site is the hybridizing site for the M13 universal sequencing primer. The Hpa/Sph subclone does not contain the coding region of the gene, but does contain the region upstream of the gene.


8062, Hind III
8057, Sph I
7540, Pst I
L11
6774, Hpa I
6388, Hind III
6230, Eco RI
MM11-Sph R
9032 bp
Pst I, Sph I
Hpa I, Sph I
7535, Hind III
L11
6774, Hpa I
6388, Hind III
6230, Eco RI
Pst/Sph
8515 bp
6779, Hind III
6388, Hind III
6230, Eco RI
Hpa/Sph
7749 bp

complementary strand sequence was reduced from about 700 nucleotides to about 400 nucleotides. Figure 3 shows the sequencing strategy and the location of the gap in complementary strand sequence data is evident. Unfortunately, for subcloning purposes, the 5' end of the gene has a large domain of repeated DNA and there are no known hexanucleotide recognition sequences available in this region with which to subclone. A further attempt was made to cover this region using the Pst I/Sph I subclone of MW11R/Sph as the substrate for deletion subcloning (25). The results were unsuccessful. Forty of these deletion subclones were screened but again they were clustered on either end of the insert. The sequence data in this region was unambiguous even though it was obtained from only one strand. Nucleotide sequence discrepancies between the original and reverse orientation clones were not observed. After comparing the sequence of L11 with other prolamin genes we found very strong homology with the aggregated gliadin B11-33 (14), the "gamma gliadin", pTag544 (33), and as expected, with pW12 [(19); Rafalski and Scheets, unpublished data] (Fig. 8).

The Nucleotide Sequence of an LMW Glutenin Gene

The nucleotide sequence obtained from the insert of the lambda clone LP11 extends 549 nucleotides upstream of

Figure 6. The nucleotide sequence of an LMW glutenin gene. The underlined sequence defines the -300 regulatory element and within this area the SV40 enhancer sequence is underlined in bold. The shaded sequence 73 nucleotides upstream of the start codon defines the TATA box, and the shaded regions downstream about 80, 140 and 150 nucleotides of the termination codon define the polyadenylation signals. The standard three letter amino acid abbreviations are positioned under each codon.

```
GCTCTGTAGGCGTCAGTTCATCTTATCATCTTAGAGGAAAATACAAAGTTAGTTTTATAAAAAGCAACCGAGTCT   75

AGAAGAACCCTCCACACGCAAGACTTCAATAATCGAGCATATCTTAACAGCCCACACACGATTGCAAACTTAGTC  150

CTACACAAGCTTTGCCTTTCTTGTTTACGGCTGACAACCTATACAAGGTTCCAAACTCGGTTGCAAAAGTGATAC  225

TATCCTGATAAGTGCGTGACATGTAAAGTTAATAAGGTGAGTCATATGTACCAAACATCGAGGTTTCTGTACTTT  300

GTGTATGATCATATGCACAACTAAAAAGCAACTTTGATGATGAATCCAAAAGTACGCTTTTGTAGCTAGTGCAAC  375

CCAACACAATGTACCAAAAAAATTCATTTCAGATGCATCCAAACAGAATTATTAAAGCCGGTGCAAAGAAGGAAA  450

AGAGGTGGTGTCCCGGCAACTATAAATAGGCATGAAGTATAAAGATCATCACAAGTACAAGCATCAAAGCCAAGC  525

AACACTAGTTAACACCAATCCACAATGAAGACCTTCCTCGTCTTTGCCCTCCTCGCTCTTGCGGGCGGCAAGTGCC  600
                    MetLysThrPheLeuValPheAlaLeuLeuAlaLeuAlaAlaSerAla   17

GTTGCGCAAATTTCACAGCAACAACAAGCACCGCCATTTTCGCAGCAACAACAACCACCATTTTCACAGCAACAA  675
ValAlaGlnIleSerGlnGlnGlnGlnAlaProProPheSerGlnGlnGlnGlnProProPheSerGlnGlnGln   42

CAACCACCATTTTCGCAGCAACAACAATCACCATTTTCGCAACAACAACAACCACCATTTGCGCAGCAACAA  750
GlnProProPheSerGlnGlnGlnGlnSerProPheSerGlnGlnGlnGlnProProPheAlaGlnGlnGln   67

CAACCACCGTTTTCACAACAACCACCAATTTCACAGCAGCAACAACCACCATTTTCACAGCAACAACAACCACA  825
GlnProProPheSerGlnGlnProProIleSerGlnGlnGlnGlnProProPheSerGlnGlnGlnProGln   92

TTTTCACAGCAACAACAACCACCATATTCGCAGCAACAACAGCCACCATATTCGCAGCAACAACAACCACCATTT  900
PheSerGlnGlnGlnGlnProProTyrSerGlnGlnGlnGlnProProTyrSerGlnGlnGlnGlnProProPhe  117

TCGCAGCAACAACAACCACCATTTTCGCAGCAACAACAACAACCACCATTTACACAGCAGCAGCAGCAACAA  975
SerGlnGlnGlnGlnProProPheSerGlnGlnGlnGlnGlnProProPheThrGlnGlnGlnGlnGlnGln  142

CAACAACAACCATTTACACAGCAACAGCAACCACCGTTTTCACAACAGCCACCAATTTCACAGCAGCAACAACCA 1050
GlnGlnGlnProPheThrGlnGlnGlnGlnProProPheSerGlnGlnGlnGlnProProIleSerGlnGlnGln  167

CCATTTTTGCAGCAACAACGACCACCATTTTCACGGCAACAACAAATACCAGTTATTCATCCATCTGTTTTGCAA 1125
ProPheLeuGlnGlnGlnArgProProPheSerArgGlnGlnGlnIleProValIleHisProSerValLeuGln  192

CAGCTAAACCCATGCAAGGTATTCCTCCAACAGCAGTGCATCCCTGTGGCAATGCAGCGATGTCTTGCTAGGTCA 1200
GlnLeuAsnProCysLysValPheLeuGlnGlnGlnCysIleProValAlaMetGlnArgCysLeuAlaArgSer  217

CAAATGTTGCAGCAGAGCATTTGCCATGTGATGCAGCAACAATGTTGCCAGCAGTTGCGGCAAATCCCCGAGCAA 1275
GlnMetLeuGlnGlnSerIleCysHisValMetGlnGlnGlnCysCysGlnGlnLeuArgGlnIleProGluGln  242

TCCCGCCATGAGTCAATCCGTGCTATCATCTACTCTATCATCCTGCAGCAGCAGCAGCAGCAACAACAACAACAA 1350
SerArgHisGluSerIleArgAlaIleIleTyrSerIleIleLeuGlnGlnGlnGlnGlnGlnGlnGlnGlnGln  267

CAACAACAACAGGGTCAGAGTATCATCCAATATCAGCAACAACAACCCCAACAGTTGGGCCAATGTGTCTCCCAA 1425
GlnGlnGlnGlnGlyGlnSerIleIleGlnTyrGlnGlnGlnGlnProGlnGlnLeuGlyGlnCysValSerGln  292

CCCCTACAGCAGTTGCAGCAGCAACTCGGGCAACAACCTCAACAACAACAATTGGCACACCAGATAGCTCAGCTT 1500
ProLeuGlnGlnLeuGlnGlnGlnLeuGlyGlnGlnProGlnGlnGlnGlnLeuAlaHisGlnIleAlaGlnLeu  317

GAGGTGATGACTTCCATTGCACTCCGTACCCTGCCAACAATGTGCAATGTCAATGTGCCGTTGTACGAAACCACC 1575
GluValMetThrSerIleAlaLeuArgThrLeuProThrMetCysAsnValAsnValProLeuTyrGluThrThr  342

ACTAGTGTGCCATTAGGCGTTGGCATCGGAGTTGGTGTCTACTGATAAGAAAAGATCTCTAGTAATATATAGTTG 1650
ThrSerValProLeuGlyValGlyIleGlyValGlyValTyr                                   356

GATCACCGTTGTTTAGTCGATGGATATGTCGATGTAGCGGTGACAAATAAAGTGTCACACAACGTCATGTGTGAC 1725

CCGCTCAAACTAGTTGTTTAAATTCTGAAATAAAATACAAATAAAGGTGTATTAAGAAAATGTTCATATCGGCAT 1800

TGTGTGGATGTCGATCTGATTGCCAT                                                   1826
```

the start codon and downstream 209 nucleotides from the termination codon (Fig. 6). The coding region is 1068 nucleotides long and starts at base number 550. This sequence codes for a 356 amino acid residue protein, but after cleavage of the signal sequence the mature protein would contain 337 residues.

Transcriptional Control Sequences

L11 has a typical TATA box (TATAAA) located about 80 nucleotides upstream of the start codon (Fig. 6). The upstream sequence of L11 shows 91% homology (458 matches of 503 nucleotides) with the published upstream sequence of LMW-Colot (15). There is also very strong homology with the upstream sequence of the barley B1-hordein gene pBHR184 (34). The transcription start site was not determined for L11 but by analogy with the published upstream sequences of the LMW glutenin gene and the B1-hordein gene, the putative transcription start would be at a position approximately 50-60 nucleotides upstream of the start codon.

There are also three consensus polyadenylation signals (AATAAA), (Fig. 6), located approximately 80, 140 and 150 nucleotides dowmstream of the stop codon.

Figure 7.

The -300 regulatory elements from S-rich prolamin genes. The underlined region highlights the SV40 enhancer sequence.

(.) indicate nucleotide matches, variant nucleotides are shown.

(*) The name applied is arbitrary; no name was given to this partial sequence in the publication.

The references for the sequences are LMW (15), pBHR184 (34), pW1020 (35), pW8233 (19), and the SV40 (36).

Class of Prolamin	Sequence	
LMW glutenin	L11	ACTATCCTGATAAGTGCGTGACATGTAAAGTTAATAAGGTGAGT
LMW glutenin	LMW*	......T.........T..........................
B1-hordein	pBHR184	....G.T.......................G..........
γ-gliadin	pW1020	G.A...TA.....T..TT...CT.........CG.....A.....
α-gliadin	pW8233	G.A...TA...T....TT....C........G......A.....
	SV40	..G....

The -300 Regulatory Element

There is a region located upstream of the coding region (Figs. 6 and 7), termed the -300 element (34), that has very strong homology (better than 95%) with the corresponding area of LMW-Colot (15) and also has better than 94% homology with the barley prolamin gene pBHR184 (34). This region upstream of the start codon is rather strongly conserved with other prolamins (Fig. 7). This indicates that there is some importance to this region, and it has been observed (9) that the core of the SV40 enhancer region (36) is almost centered in the -300 element (34), perhaps indicating a regulatory role for this region. It is also interesting to note that the seven nucleotides homologous to the SV40 enhancer region on the prolamin genes shown in Fig. 7 are conserved even though the nucleotides immediately upstream and downstream show considerable variation.

The enhancer region in SV40 stimulates the rate of transcription from eucaryotic promoters so by analogy this -300 region in prolamin genes may aid in the amplification of the storage proteins during their critical period of production. Kreis et al. have postulated that gene expression is controlled, at least in part, by sequence-specific DNA-binding proteins that recognize short sequences in the 5' flanking region of prolamin genes (9).

Obviously the -300 element is a prime candidate for this role, due to its conserved nature. In L11, the 5' end of the -300 element is at a position 327 nucleotides upstream of the start codon (Fig. 6).

There has been some recent evidence to support the hypothesis that the genes encoding the wheat storage proteins are regulated in a tissue-specific fashion (15). A chimeric gene was formed by fusing the gene encoding cloramphenical acetyl transferase (CAT) to the upstream sequence of a putative LMW glutenin gene. In order to pinpoint the important region in the upstream sequence of the LMW glutenin gene, a series of deleted upstream regions were fused to the functional CAT gene. By assaying CAT activity the authors were able to show that the essential region was within 326 nucleotides upstream relative to the transcription start site. A fused gene containing only 160 upstream nucleotides did not show any CAT activity, indicating that the enhancer element had been deleted. Activity of the chimeric gene product was found exclusively in the seeds of the transgenic tobacco plants and in no other tissues of the plants. This is the first direct evidence of tissue-specific regulation of a prolamin gene (15). This work strongly supports the assertions of Forde *et al.* (34) that the -300 element plays a key role in the regulation of the prolamin genes. This analogous region of L11 is strongly homologous with

that of the LMW glutenin upstream sequence (15), which would suggest that it might behave in the same way.

Signal Sequence

L11 has a putative signal sequence of 19 amino acid residues at the amino terminus of the protein. The signal sequence has a net charge of +1, due to a sole lysine residue adjacent to the amino terminus, and in this regard is identical to many eukaryotic signal sequences studied (37). The positively charged lysine residue near the amino terminus is followed by 17 hydrophobic residues which are contiguous up to the putative cleavage site between residues 19 and 20. In comparison to the signal sequence of wgb11-33 (14) the first nine amino acid residues are identical and of the remaining ten amino acid residues in the signal sequence only four are identical. The signal sequence cleavage site is putative, and is estimated by comparing L11 with B11-33 (14).

Amino Acid Composition

The amino acid composition of L11 is very unusual in comparison with many other proteins, yet strongly homologous with that of other S-rich prolamin genes, (Table 1). It is very rich in proline and glutamine,

Table I Amino Acid Compositions of S-rich Prolamins. The amino acid compositions of L11 and B11-33 (14) were computed without the signal sequence using deduced amino acid sequences. The LMW glutenin and aggregated gliadin from the cultivar CWW are from amino acid analysis of exracted protein fractions (15). The α and β-gliadin are from Bietz et al. (16).
Glx denotes glutamine and gutamic acid, Asx denotes asparagine and aspartic acid.

TABLE I Amino Acid Compositions of S-rich Prolamins (mol%)

type	LMW glu	LMW glu	agg.gli	agg.gli	α-gli	β-gli
sequence	L11	CWW	CWW	B11-33	α_8	β_5
Amino Acid						
Asx	0.89	1.45	1.82	0.70	3.0	2.6
Thr	2.37	2.26	2.36	2.80	1.6	1.7
Ser	8.01	6.47	6.04	9.82	5.2	5.3
Glx	41.84	40.29	38.31	36.41	37.2	39.6
Pro	13.65	15.74	15.91	11.57	15.5	15.8
Gly	2.08	3.27	3.37	3.51	2.5	1.9
Ala	2.37	2.21	2.72	3.16	2.9	3.5
Cys	2.37	2.13	2.62	3.16	1.9	1.8
Val	3.86	4.28	4.38	5.96	4.0	4.0
Met	1.48	1.02	1.02	1.75	1.2	1.1
Ile	5.34	4.29	4.16	4.21	4.1	3.8
Leu	5.34	6.87	7.35	7.72	8.1	7.9
Tyr	1.78	0.91	1.14	1.40	3.1	2.5
Phe	4.75	4.62	4.77	3.85	3.9	3.7
His	1.19	1.65	1.37	1.40	2.5	2.3
Lys	0.30	0.63	0.67	0.35	0.5	0.2
Arg	2.37	1.90	2.01	2.10	2.4	2.3

hence the name " prolamin ", and is rich in the sulfur-containing amino acids. Its classification as an LMW glutenin is partly based on the assertion that it has a slightly higher glutamime and glutamate concentration and a somewhat lower occurence of cysteine relative to the aggregated gliadin (12). As mentioned earlier, wheat protein is low in lysine, and L11 reflects that trait with only one lysine residue in the mature protein. Lysine is the most notably deficient amino acid, while threonine and valine are the other essential amino acids that are somewhat limiting. Another characteristic of the S-rich prolamins is that they are very low in charged residues, with the lysine deficiency already noted, L11 only has 8 arginine residues, 4 histidine residues, 4 glutamic acid residues, and no aspartic acid residues for a maximum of 17 charged residues in the mature 337 amino acid residue protein.

The S-rich prolamins can be divided at the amino acid level into proline-rich and proline-poor domains (9). The proline rich domain is mostly composed of repeated peptides. The repeat found in L11 is an octapeptide with the sequence QQQQPPFS, using the standard one letter abbreviations. This consensus repeat appears eighteen times within the coding sequence of L11 and is obviously related to the repeat found in the aggregated gliadin B11-33 (14) which has the repeat sequence PQQPPFS.

Discussion

Homologous Regions of the Proline-Poor Domain

There are three regions within the proline-poor domain which show strong homology at the amino acid level. (Fig 8). Kreis et al. termed these regions A, B, and C, with the intervening regions being termed I_1-I_4 (9). There is very strong homology among regions A, B, and C between various S-rich prolamins and the LMW glutenin gene L11. A notable feature of these regions is that they are all relatively rich in cysteine and charged residues which would effect the structure and interactions of these proteins with other wheat storage proteins. This may be the reason that these regions are conserved throughout the entire family of S-rich prolamins.

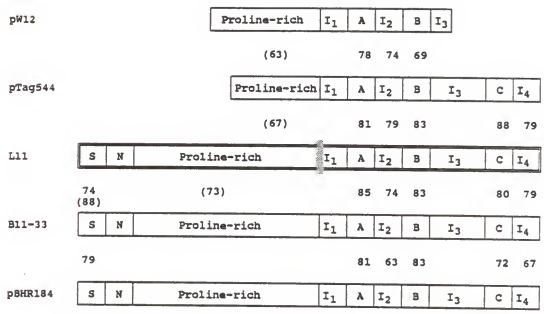
Homology at the Nucleotide Level

We have found very strong homology with other prolamin genes with the aid of the alignment program **Nucaln** of Wilbur and Lipman (38). There is very strong homology between the closely related aggregated gliadin sequence of B11-33 (14) and the sequence of L11, Fig. 9, and there is also strong homology between L11 and the sequences of pTag544 (34) and pW12 (19); Rafalski and

Figure 8A. Homology between L11 and other prolamin genes. The vertical bar in L11 divides the proline-rich and proline-poor domains. Homology at the amino acid level between L11 and each of the other sequences is shown as a percentage of identical residues. Identities at the nucleotide level are shown parenthetically and were generated by the results given in Fig. 9. The various boxed areas (region I_3 for example) are not drawn to scale.

B. The amino acid positions assigned to the various regions shown in Fig. 8A. The data for B11-33 and pBHR184 was taken directly from Kreis et al. (9), with the additional references; B11-33 (14), pBHR184 (34), pTag 544 (33) and pW12 (19); Rafalski and Scheets, unpublished data.

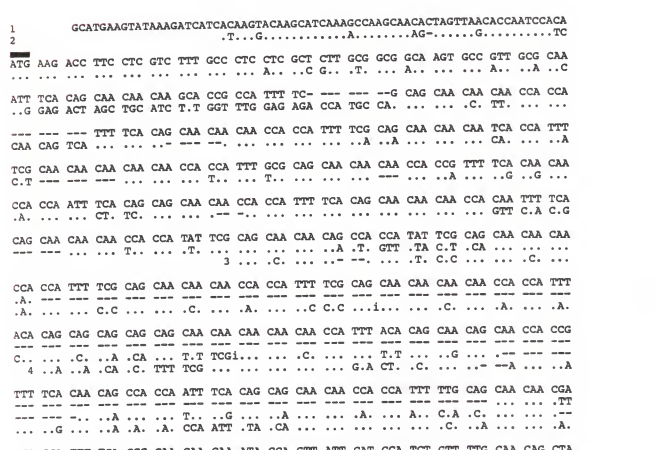
A



B

	L11	B11-33	pBHR184	pTag544	pW12
Signal	1-19	1-19	1-19	n.d.	n.d.
N-terminal	20-22	20-33	—	n.d.	n.d.
Proline-rich	23-176	34-108	20-98	1-43	1-79
I_1	177-182	109-119	99-128	44-46	80-82
A	183-209	120-146	129-155	47-73	83-109
I_2	210-228	147-165	156-174	74-92	110-128
B	229-263	166-200	175-209	93-128	129-163
I_3	264-317	201-265	210-259	129-180	164-176
C	318-342	266-290	260-284	181-205	n.d.
I_4	343-356	291-304	285-293	206-219	n.d.

Figure 9. The compiled results computed from the Nucaln program. Sequence 1 is a portion from base 480 (using the numbering from Fig. 6) to the end of the sequence of the LMW glutenin sequenced in this study. Sequence 2 is the entire sequence of B11-33 (14). Sequence 3 is the entire sequence from the cDNA clone pW12 (19); Raflaski and Scheets, unpublished data. Sequence 4 is the entire sequence from the cDNA clone pTag544 (33). The dots (.), indicate identical nucleotides (relative to L11), the dashes (-), indicate gaps introduced by the computer program to maximize homology, the (i) indicates sequence insertions in a sequence, relative to L11, and the bar over the first triplet codon indicates the start codon of L11.


1 GCATGAAGTATAAAGATCATCACAAGTACAAGCATCAAAGCCAAGCAACACTAGTTAACACCAATCCACA
2 .T...G.....A.....AG-.....G.....TC
ATG AAG ACC TTC CTC GTC TTT GCC CTC CTC GCT CTT GCG GCG GCA AGT GCC GTT GCG CAA
ATT TCA CAG CAA CAA CAA GCA CCG CCA TTT TC- --- --- --G CAG CAA CAA CAA CCA CCA
..G GAG ACT AGC TGC ATC T.T GGT TTG GAG AGA CCA TGC CA. ... ... .C. TT. ... ...
--- --- --- TTT TCA CAG CAA CAA CAA CCA CCA TTT TCG CAG CAA CAA CAA TCA CCA TTT
CAA CAG TCA ... .- --- --. ... ...A ..A ... ... CA. ... ..A
TCG CAA CAA CAA CAA CAA CCA CCA TTT GCG CAG CAA CAA CAA CCA CCG TTT TCA CAA CAA
C.T --- --- --- ... ... T.. ... T.. ... ... --- ... ..A ... ..G ..G ...
CCA CCA ATT TCA CAG CAG CAA CAA CCA CCA TTT TCA CAG CAA CAA CCA CAA TTT TCA
.A. ... ... CT. TC. ... .-- -.. ... ... GTT C.A C.G
CAG CAA CAA CAA CCA CCA TAT TCG CAG CAA CAA CAG CCA CCA TAT TCG CAG CAA CAA CAA
--- --- ... ... T.. ... .T. ... ...A .T. GTT .TA C.T .CA ... ...
3 ... .C. ... ..- --. ... .T. C.C ... ... .C. ...
CCA CCA TTT TCG CAG CAA CAA CAA CCA CCA TTT TCG CAG CAA CAA CAA CCA CCA TTT
.A. --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
.A. ... ... C.C ... ... .C. ... .A. ... ..C C.C ...i... ... .C. ... .A. ... .A.
ACA CAG CAG CAG CAG CAG CAA CAA CAA CAA CCA TTT ACA CAG CAA CAG CAA CCA CCG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C.. ... .C. ..A .CA ... T.T TCGi... ... .C. ... ... T.T ... ..G ... .-- --- ---
4 ..A ..A .CA .C. TTT TCG ... ... G.A CT. .C. ... ..- --A ... ..A
TTT TCA CAA CAG CCA CCA ATT TCA CAG CAG CAA CAA CCA CCA TTT TTG CAG CAA CAA CGA
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ... ... .TT
--- --- -.. ..A ... ... T.. ..G ... ..A ... .A. ... A.. C.A .C. ... .--
... ..G ... ..A .A. .A. CCA ATT .TA .CA ... ... .C. ..A ... ... .A.
CCA CCA TTT TCA CGG CAA CAA CAA ATA CCA GTT ATT CAT CCA TCT GTT TTG CAA CAG CTA
GT- --- --- --- --. ... ..G ... ..C ..T A.. G.. ..G ... ..C ... .G ... ...
-.. ... ... .Ai ... ... TC. G.. ..G ... ... A.C ... ..G ... ...
.A. ... G.. CT. .CA ... ...T. T.. G.. ... ... A.C ... ..G ... ...
AAC CCA TGC AAG GTA TTC CTC CAA CAG CAG TGC ATC CCT GTG GCA ATG CAG CGA TGT CTT
... ...A ... ...G ... ...G. ... ..A ... ... .CA .A. C.. ...
C.. ... ...G ... ..A ... ..G. ... ... .CA .A. A.. ...
... ...G ... ..A ... .G. ... ... .CA .A. A.. ...
GCT AGG TCA CAA ATG TTG CAG CAG AGC ATT TGC CAT GTG ATG CAG CAA CAA TGT TGC CAG
... ...G ... .G. ... ...G. ... ...A ... ...
... .G ... ... TG. ... ..T .G. ... ... .GC ... ... .C.
... .G ... ... .G. ... ...A ... ...

Scheets, unpublished data.

Summary

The goal of this study was to characterize a gene from the group of wheat storage proteins called the low molecular weight glutenins. The wheat storage protein genes encoding the gliadins and the high molecular weight glutenins have already been well characterized (9). The LMW glutenin subunits have not been studied in detail at the nucleotide level, and this lack of data is recognized as an omission of important information concerning an important class of wheat storage proteins (9). An important result of our work is that since the nucleotide sequence of the LMW glutenin gene, L11, was derived from a genomic clone and not a cDNA clone, as has often been the case in the past, it contains the intact sequences upstream and downstream of the gene. The lambda clone we sequenced, LP11, contained the upstream regulatory region. This upstream "-300 element" (34) has been shown to confer tissue-specific regulation when introduced into tobacco plants via A. tumefaciens mediated transformation (12). Comparison of the -300 element from the sequence determined in this study and the -300 element reported by Colot et al. (15) shows better than 95% homology between the two regions.

It has been observed that the α- and γ-gliadins from wheat, the B1-hordeins from barley, and genes from the multi-gene family of maize all have the -300 element (34). The proposed regulatory role of the -300 element is now more tenable due to the results of Colot et al. (15). As noted in the results section, the region within the -300 element homologous to the SV40 enhancer region is perfectly conserved among the prolamin genes even though the surrounding sequence shows considerable variation (Fig. 7). This suggests that it may be a control element within the -300 regulatory element. In order to test this hypothesis one could make use of site-directed mutagenesis utilizing a set of degenerate oligonucleotides (40), thereby causing point mutations at each of the seven SV40 enhancer region nucleotides. Once this was completed the aforementioned transformation system could be employed to assay the effect that various mutations have on tissue-specific regulation.

A project already underway in our laboratory is to screen wheat genomic DNA with a probe containing the upstream regulatory region from pW1020, a γ-gliadin genomic clone (35), to see if any genes outside the storage protein gene family can be identified which may contain a related regulatory element. If a gene outside the storage protein realm were found, it would be sequenced to determine what similarities to the prolamin -

300 element arose.

As our knowledge of the regulation of the prolamin genes broadens, the practical and important goal of increasing the lysine content of wheat may be achieved. This may be possible by adding codons for lysine into several genes encoding prolamins. Since there are a large number of these genes due to the hexaploid genome, the goal may be achieved without the accompanying decrease in yield associated with low prolamin varieties of barley (4), or negatively effecting baking properties because a vast majority of the storage proteins could be left in their native state.

References

1. Market Report, (8 Dec.,1987) International Wheat Council, London
2. Shewry, P.R., Miflin, B.J., and Kasarda, D.D. (1984) Phil. Trans. R. Soc. Lond. B **304**, 297-308
3. Payne, P.I., and Rhodes, E.M., (1982) Encyclopedia of Plant Physiology **14A**, 346-369
4. Bright, S., Shewry, P.R., (1983) CRC Crit. Rev. Pl. Sci. 1, 49-93
5. Hoseney, R.C. (1986) Principles of Cereal Science and Technology. American Association of Cereal Chemists St. Paul, Minnesota
6. Rafalski, J.A. (1986) Gene 43, 221-229
7. Mecham, D.K., Fullington, J.G., and Greene, F.C. (1981) J. Sci. Food Agric. 32, 773-780
8. Payne, P.I., Holt, L.M., Lawrence, G.J., and Law, C.N. (1982) Qual. Plant. Plant Foods Hum. Nutr. 31, 229-241
9. Kreis, M., Shewry, P.R., Forde, B.G., Forde, J., and Miflin, B.J. (1985) Oxford Surv. Plant Mol. Cell Biol. V. 2, 253-317
10. Miflin, B.J., Burgess, S.R., and Shewry, P.R. (1981) J. Exp. Botany 32, 199-219
11. Osborne, T.B. (1924) The Vegetable Proteins, Longmans, Green & Co.

12. Tatham, A.S., Field, J.M., Smith, S.J., and Shewry, P.R. (1987) J. Cereal Sci. 5, 203-214

13. Woychik, J.H., Boundy, J.A., and Dimler, R.J. (1961) Arch. Bioch. Biophys. 94, 477-482

14. Okita, T.W., Cheesbrough, V., and Reeves, C.D. (1985) J. Biol. Chem. 260, 8203-8213

15. Colot, V., Robert, L.S., Kavanagh, T.A., Bevan, M.W., and Thompson, R.D. EMBO J. 6, 3559-3564

16. Bietz, J.A., Huebner, F.R., Sanderson, J.E., and Wall, J.S. (1977) Cereal Chem. 54, 1070-1083

17. Loenen, W.A.M. and Blattner, F.R. (1983) Gene 26, 171-179

18. Murray, M.G., Kennard, W.C., Drong, R.F., and Slightom, J.L. (1984) Gene 30, 237-240

19. Rafalski, J.A., Scheets, K.M., Metzler, M., Peterson, D.M., Hedgcoth, C., and Soll, D. (1984) EMBO J. 3, 1409-1415

20. Scheets, K., Rafalski, J.A., Hedcoth, C. and Soll, D.G. (1985) Plant Sci. Lett. 37, 221-225

21. Forde, J. Malpica, J-M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C., and Miflin, B.J. (1985) Nucleic Acids Res. 13, 6817-6832

22. Thompson, R.D., Bartels, D., and Harberd, N.P. (1985) Nucleic Acids Res. 13, 6833-6846

23. Sugiyama, T., Rafalski, A., Peterson, D. and Soll, D.

(1985) Nucleic Acids Res. 13, 8729-8737

24. Shewry, P.R., Miflin, B.J., Lew, E.J., and Kasarda, D.D. (1983) J. Exp. Botany 34, 1403-1410
25. Dale, R.M.K., McClure, B.A., and Houchins, J.P. (1985) Plasmid 13, 31-40
26. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467
27. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning, a Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York
28. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517
29. Cohen, S.C., Chang, A.C.Y., Boyer, H.W., and Helling, R.B. (1973) Proc. Natl. Acad. Sci. U.S.A. 70, 3240-3244
30. Rusche, J.R., Howard-Flanders, P. (1985) Nucleic Acid Res. 13 1997-2008
31. Messing, J. (1983) Methods Enzymol. 101, 20-78
32. Hanahan, D. (1983) J. Mol. Biol. 166, 557-580
33. Bartels, D., and Thompson, R.D. (1983) Nucl. Acids Res. 11, 2961-2977
34. Forde, B.G., Heyworth, A., Pywell, J., and Kreis, M. (1985) Nucleic Acids Res. 13, 7327-7339
35. Scheets, K.M., manuscript submitted.
36. Weiher, H., Konig, M., and Gruss, P. (1983) Science 219, 626-631
37. Heijne, G. (1984) EMBO J. 3, 2315-2318

38. Wilbur, W.J., and Lipman, D.J. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 726-730

39. Payne, P.I., Holt, L.M., Jackson, E.A., and Law, C.N. (1984) Philos. Trans. R. Soc. Lond. B Biol. Sci. 304, 359-371

40. Hill, D.E., Oliphant, A.R., and Struhl, K. (1987) Methods Enzymol. 155, 558-568

THE NUCLEOTIDE SEQUENCE OF A GENE ENCODING A LOW MOLECULAR WEIGHT GLUTENIN SUBUNIT FROM HEXAPLOID WHEAT

by

ERNEST GERARD PITTS

B.S. St. John's University, 1982

AN ABSTRACT OF A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

GRADUATE BIOCHEMISTRY GROUP

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1988

Wheat is the most produced and consumed cereal grain in the world. The storage proteins of wheat are the largest group of proteins in wheat and therefore are the subject of intensive study. All of the classes of wheat storage proteins have been well characterized at the nucleotide level with the exception of two groups: the aggregated gliadins and the low molecular weight (LMW) glutenin subunits. We have determined the nucleotide sequence of a gene encoding a LMW glutenin subunit which was screened and cloned from a wheat genomic library constructed in lambda Charon 32.

Within the nucleotide sequence a 1068 nucleotide open reading frame was found which did not contain introns. Five-hundred and forty nine nucleotides upstream of the gene and 209 nucleotides downstream of the gene were also sequenced. The resulting protein contained 356 amino acid residues and its composition was typical of the prolamins; high glutamine and proline content, and a low occurence of lysine and other charged residues. The predicted molecular weight of the protein is 41 kDa.

Typical control elements upstream (TATA box) and downstream (polyadenylation signals) were found. In addition, the upstream -300 regulatory element found was in excellent agreement with several of the -300 regions previously reported, suggesting that the gene is transcriptionally active.